Technical Document 2721
October 1994

Vocoded KING Data Base

P. C. Grossnickle

EXECUTIVE SUMMARY

OBJECTIVE

Develop a data base of vocoded speech; that is, speech which has been passed through several popular voice coders (vocoders) and their corresponding decoders.

RESULTS

The KING data base consists of samples from 26 male speakers, each about 50 s long, recorded sequentially on DAT tape. We produced two products:

(1) Five DAT tapes. Each tape contains the original KING data base on one channel and the vocoded version on the other.

(2) Raw binary files, one for each vocoder and speaker combination.

We fell short of our original goal of testing Rome Laboratory speaker ID software against samples of vocoded speech.

RECOMMENDATIONS

Two tasks are recommended as a continuation of this effort. First, transfer Rome Laboratory speaker recognition software to a platform at the Naval Command, Control and Ocean Surveillance Center (NCCOSC) RDT&E Division. And second, evaluate the performance of this software against the original KING data base and against vocoded versions of the same data base. Thus, one could evaluate the degradation of this software when presented with samples of vocoded speech.

CONTENTS

Figures

Table

INTRODUCTION

Jim Cupples of Rome Laboratory conceived this project to test some current speaker ID software on samples of vocoded speech; i.e., speech that has been passed through a voice coder (vocoder) and decoder. As the project evolved, we focused our efforts on establishing a valid vocoded speech data base, which this report describes. We suggest the testing envisioned by Jim Cupples, as a follow-on effort.

Rome Laboratory provided a copy of the KING data base (Session 1) recorded on a Digital Audio Tape (DAT) at 48K (samples per second). This tape contains three parts, each about 22 minutes long:

1. the original KING data base (Session 1), 22 speakers
2. (1) corrupted with noise at 18 dB SNR
3. (1) corrupted with noise at 10 dB SNR

All data products described here spring from this tape.

TOOLS USED

We employed two commercial products in this work:

1. Piranha 3111 Evaluation Board by DSP Research, Sunnyvale, CA
2. Sound Stage system from Turtle Beach Systems, York, PA

Appendix A describes the Piranha 3111 board in detail. This board is IBM PC compatible. It accepts analog speech as input, processes this signal through software selectable vocoders, and generates an analog output.

The performance of both the vocoder used and its associated analog circuitry is fundamental to this work. Appendix A also details our evaluation of the Piranha 3111. It describes measurement of maximum input levels, SNR, and distortion products. Appendix A also describes the vendor's validation of the vocoder software.

The Sound Stage system consists of a interface box, an IBM-compatible card, and software. This system takes digital audio from a DAT drive and permits editing and resampling of the resulting files.

DATA PRODUCTS

Two products resulted from this effort:

1. Five DAT tapes. Each tape contains, on channel R, the original King base and the same corrupted with noise at 18 and 12 SNR. Channel L contains the vocoded version of channel R. For synchronization, each track starts with a 0. 5-s, 1. 0-kHz tone burst.
2. 156 raw binary files. There are six sets of files; one for the original and one for each of five vocoders. Each set consists of 26 40-s files, one for each speaker. Each file results from resampling to 8K sps at 16-bits per sample. Thus each file contains

   8-K samples/s x 2 bytes/sample x 40 s = 640-K bytes.

These files stem only from the clean KING data. We ignored versions of the KING data corrupted by noise. We can offer this data on a 150-Mbyte Bernoulli disk for IBM PC-compatible machines or on a DAT tape in TAR format for UNIX based computers.

TAPE GENERATION

RECORD SETUP

Figure 1 shows generation of the DAT tapes. As indicated, we played the original KING tape on a DAT drive and sent the analog output to both the vocoder and the channel R analog input of a second drive. Simultaneously, we recorded the vocoder analog output on channel L of the same second drive. Thus, on the product tapes, the original KING signal (channel R) suffers the same degradation as the vocoded version; they both are processed by the D/A and A/D converters of the respective tape drives.

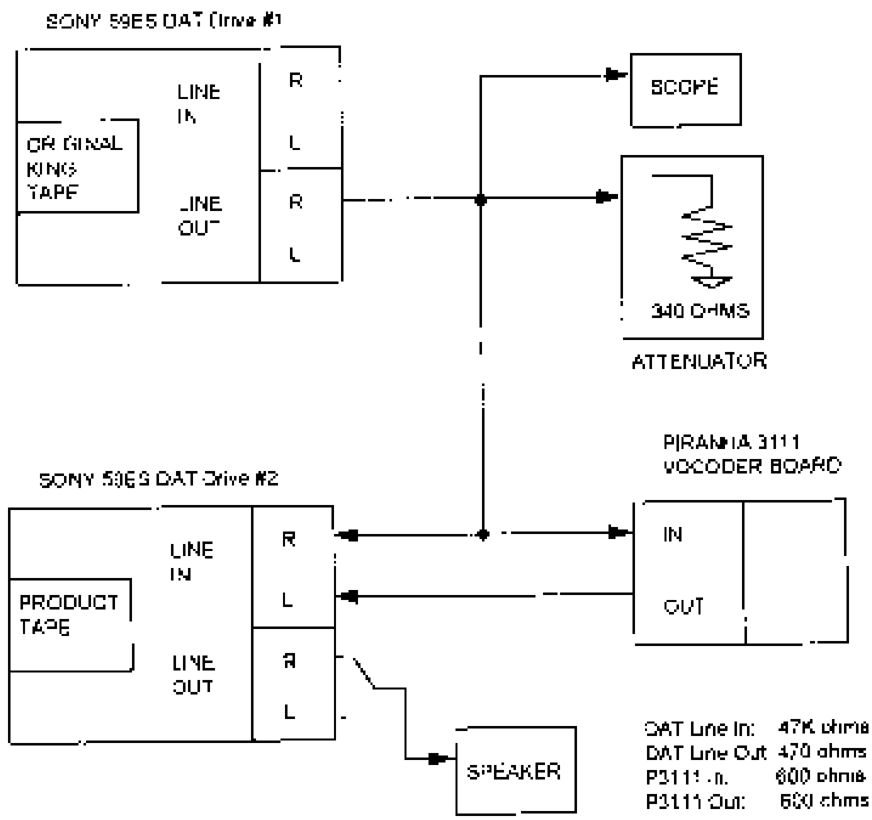


Figure 1. Generation of product tapes.

Both the original and product DAT tapes were recorded at 48K sps.

RECORD LEVELS

Two levels are important here. We must not overdrive the A/D converters of either the vocoder or the recording DAT drive.

The testing in appendix A shows that the vocoder input becomes over-driven at about 3 Vp-p (volts peak-to-peak). We inserted an attenuator before the vocoder input to ensure that this level will not exceed about 2 Vp-p for the loudest of the 26 speakers.

The Sony 59ES DAT drives feature the usual record level control knobs. We set these record levels at 3.5 and 4.5 (units marked

on control knob) for channels L and R. This equalized the record levels between channels and prevented saturation by the loudest speaker, as indicated by record level meter. At these adjustments, the average power level of some speakers was down by about 10 dB.

The 10-dB power level variation noted above means some loss in SNR for the weaker speakers. However, the vocoder and the DAT drive input A/D converters have 14 and 16 bits, respectively. So this SNR loss is not significant.

INDEXING

As an index mark, we inserted a 1-kHz, 500 ms tone burst at the start of the original tape. We intended for this mark to provide a reference for aligning original and vocoded speech segments. However, for some vocoders, the tone location is blurred by its rise time. So this mark was not as useful as originally hoped.

PRODUCTS

Table 1 listed the five product tapes. On each tape, channel R contains the vocoder input and channel L has the vocoder output.

Table 1. DAT tapes.

Tape	Vocoder	Standard	Rate (bps)
1	LPC–10e	USFS 1015	2400
2	CELP	USFS 1016	4800
3	VSELP	TIA IS–54	8000
4	LD–CELP	ITU G.728	16000
5	ADPCM	ITU G.726	32000

Note: channel R – vocoder input
channel L – vocoder output

SPEAKER FILES

Figure 2 details the generation of individual files for each speaker and vocoder combination.

Using Sound Stage, we resampled the product tapes from 48K sps to 24K sps and saved these as disk files. The following tape tracks were processed:

1. Channel L (vocoded version) of each tape in table 1.
2. Channel R (original) of tape 1.

We processed only the first part (about 22 minutes) of each tape track corresponding to the original KING data base uncorrupted by noise.

The resulting six files each correspond to about 22 minutes of speech and are quite big, about 65 Mbytes.

Next we segmented each of the above files into 26 segments (one per speaker) each 40 s long, resampled to 8K sps, and wrote these to disk as 16-bit raw binary files.

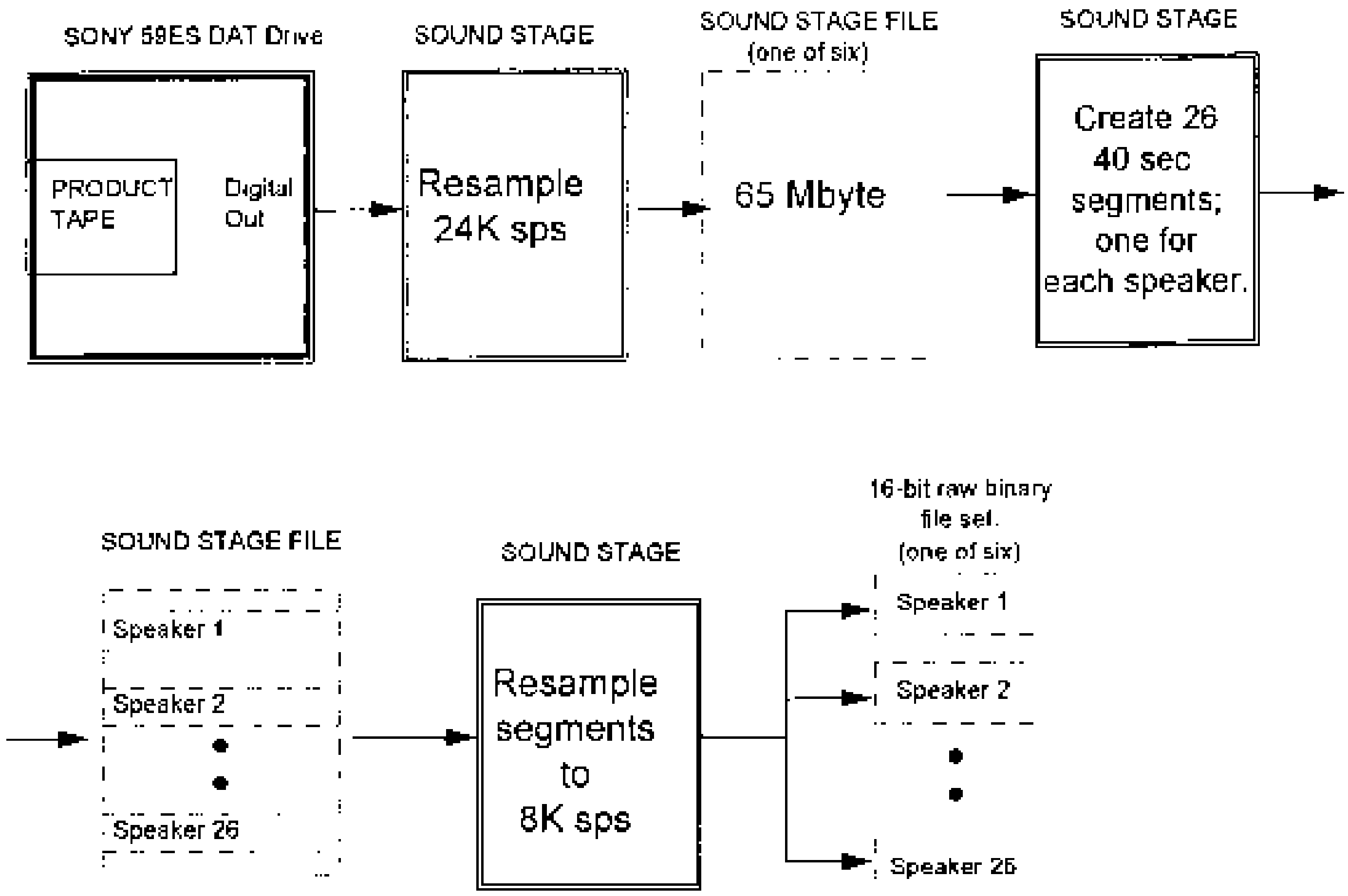


Figure 2. Generation of raw binary speaker files.

SPEAKER SEGMENT ALIGNMENT

Vocoders introduce time delay. For each speaker, there are six files, the original and five vocoded versions. We wanted alignment in time of these six files. Such alignment will allow comparisons in speaker ID and other work. The problem then is to select a feature in the recorded data on which to key the start of each 40 s segment.

Figure 3 shows a cartoon time plot of the recorded data. Note that between speakers, the noise level is dramatically reduced. Then there is a distinct jump in noise level, presumably at recorder turn-on. This is followed by the initiation of the speech waveform. This pattern is common to both the original and vocoded sound files; it is particularly evident when viewing a time plot with an expanded amplitude scale.

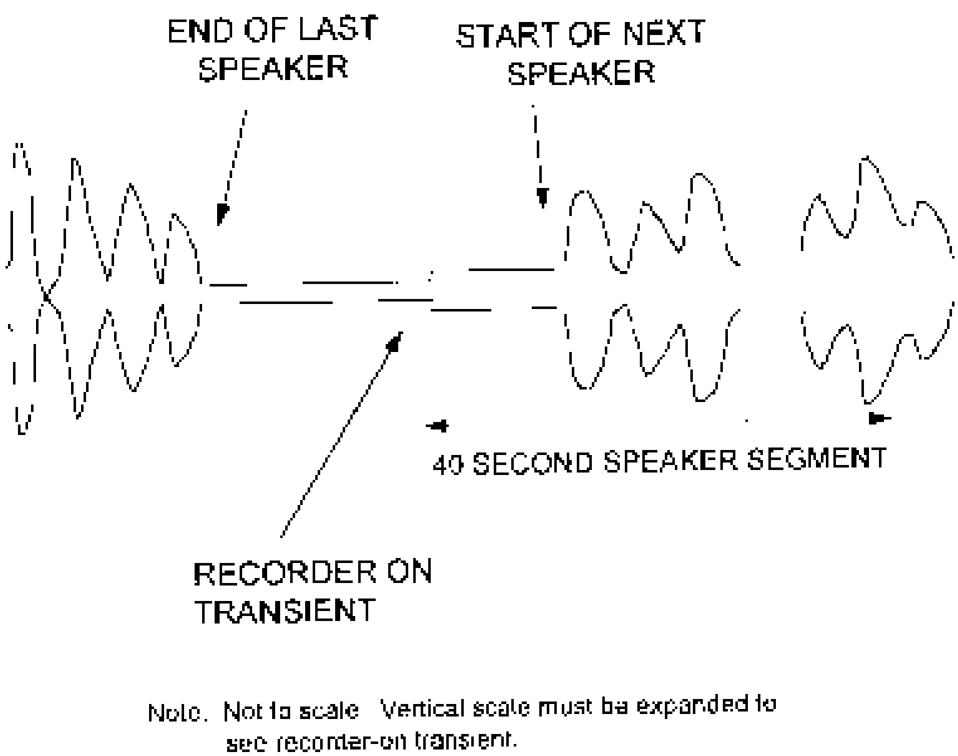


Figure 3. Selection of 40 second speech segment.

We chose the jump in noise at the start of each monologue to define the start of each 40 s speaker file. By expanding the time scale in Sound Stage, this starting point can be selected. The error in placing this starting point is estimated to be from 1 to 10 ms, depending on the vocoder.

RELATED WORK

Reference 1 describes a related project. Here we attempted speaker ID using parameters from the LPC-10 and Code Excited Linear Prediction (CELP) vocoders. The input to these vocoders was the clean KING data base. We wrote software to intercept the serial bit stream from each of the vocoders and recorded the transmitted parameters.

We can provide the parameter data base described above. For both LPC-10 and CELP, there are 26 parameter files, one for each speaker, corresponding to 36 s of speech.

SUMMARY AND RECOMMENDATIONS

We offer here a vocoded speech data base derived from the KING data base of 26 speakers. Five popular vocoders are represented. The vocoded data base is available in two forms: DAT tapes and individual speaker sound files.

We recommend that the speaker files be processed by speaker recognition software to measure the degradation caused by vocoding.

REFERENCE

1. Grossnickle, P.C. 1994. “Speaker Recognition using Vocoded Speech Parameters,” NCCOSC Technical Document 2722 (October). Naval Command, Control and Ocean Surveillance Center, RDT&E Division, San Diego, CA 92152-5000.

APPENDIX A

PIRANHA 3111 VOCODER EVALUATION

INTRODUCTION

DSP Research, Inc. , of Sunnyvale CA, markets the Piranha 3111, a module designed for audio signal processing. The P3111 consists of a TI (Texas Instruments) TMS320C31 processor, an analog interface chip, and other integrated circuits (figure A-1).

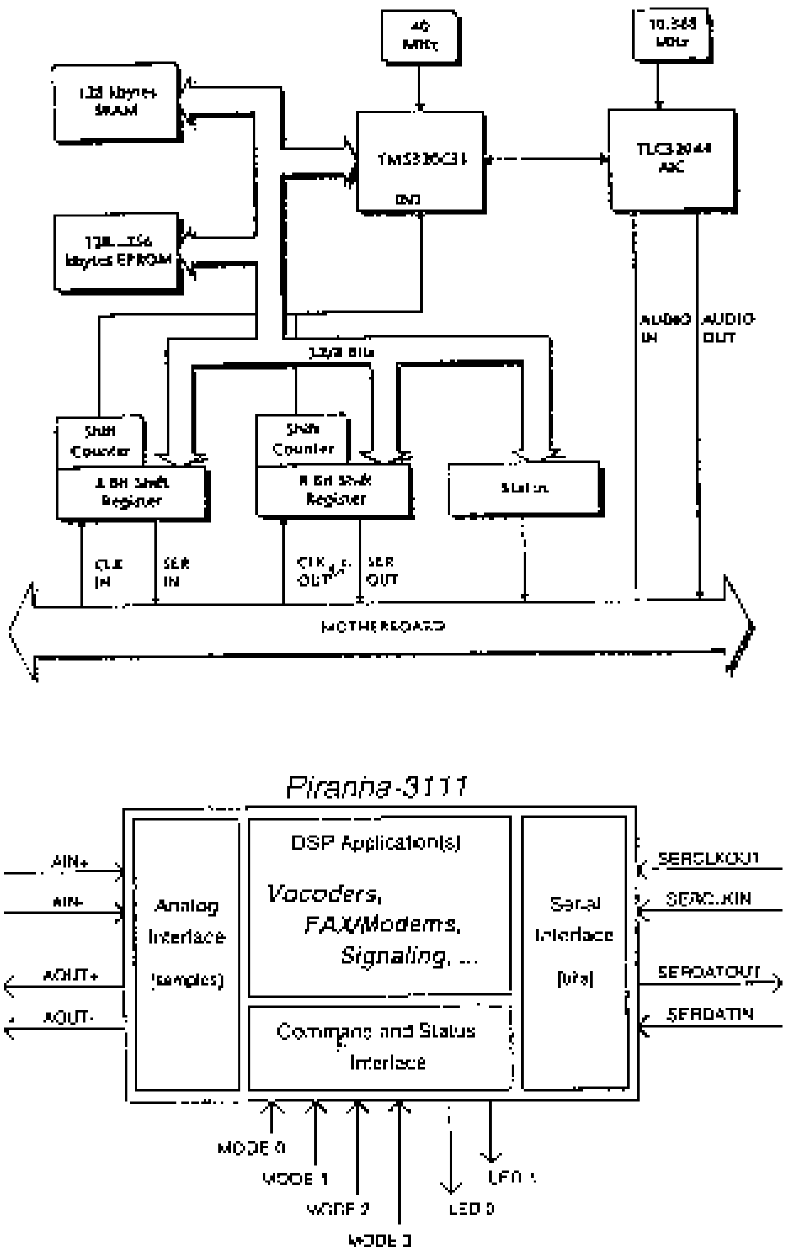


Figure A-1. Piranha 3111 block diagram (from P3111 reference manual).

DSP Research also sells the P3111 Evaluation Board for use in an IBM AT compatible PC. This board has a P3111 module, and audio and host PC interface circuitry (figure A-2).

The supplied PROM on the P3111 module contains TMS320C31 code for several popular voice coders or vocoders (table A-1).

Table A-1. Piranha 3111 vocoder package.

Vocoder	Rate (bps)
USFS 1015 LPC10e	2400
USFS 1016 CELP USFS 1016 CELP	4800 7200
TIA IS-54 VSELP	8000
ITU G.728 LD-CELP ITU G.728 LD-CELP ITU G.728 LD-CELP	12000 14400 16000
ITU G.726 ADPCM ITU G.726 ADPCM ITU G.726 ADPCM	24K 32K 40K
ITU G.711 u-law ITU G.711 u-law ITU G.711 u-law	48K 56K 64K
ITU G.722 SB-ADPCM ITU G.722 SB-ADPCM ITU G.722 SB-ADPCM	48K 56K 64K
DSPSE-CVSD	16K

With the supplied PROM, the P3l11 digitizes an input analog speech signal and encodes this as a serial bit-stream, as per the vocoder selected. The P31111 can be configured to immediately decode this bit-stream and convert the result to an output analog signal. DSP Research supplies software to run on the host PC and control the P3111 Evaluation Board. Thus, a user can select a particular vocoder, input analog speech to the evaluation board, and observe the decoded analog output (vocoded speech). Software to capture the serial bit-stream on a PC is available from the author.

OBJECTIVES

One goal is to evaluate the analog circuitry of the P3111 and its evaluation board. This task is simplified by a loop-back PROM available from DSP Research. This PROM causes the TMS320C31 processor to output its input data without modification. Thus, the A/D and D/A converters are essentially back-to-back, allowing the entire analog path to be characterized.

TI's Voice Band Analog Interface chip, the TLC320044, provides the A/D and D/A circuits and associated filters. The converters are 14-bit devices. The only other analog devices involved are the 600 ohm isolation transformers at the evaluation board I/O ports.

These tests focus on spectrograms of the P3111 analog output, given a single tone input and with the loop-back PROM installed. Since we did not have access to the A/D output data, no static tests were conducted. However, spectrograms can reveal non- linear effects such as frozen bits or differential nonlinearity.

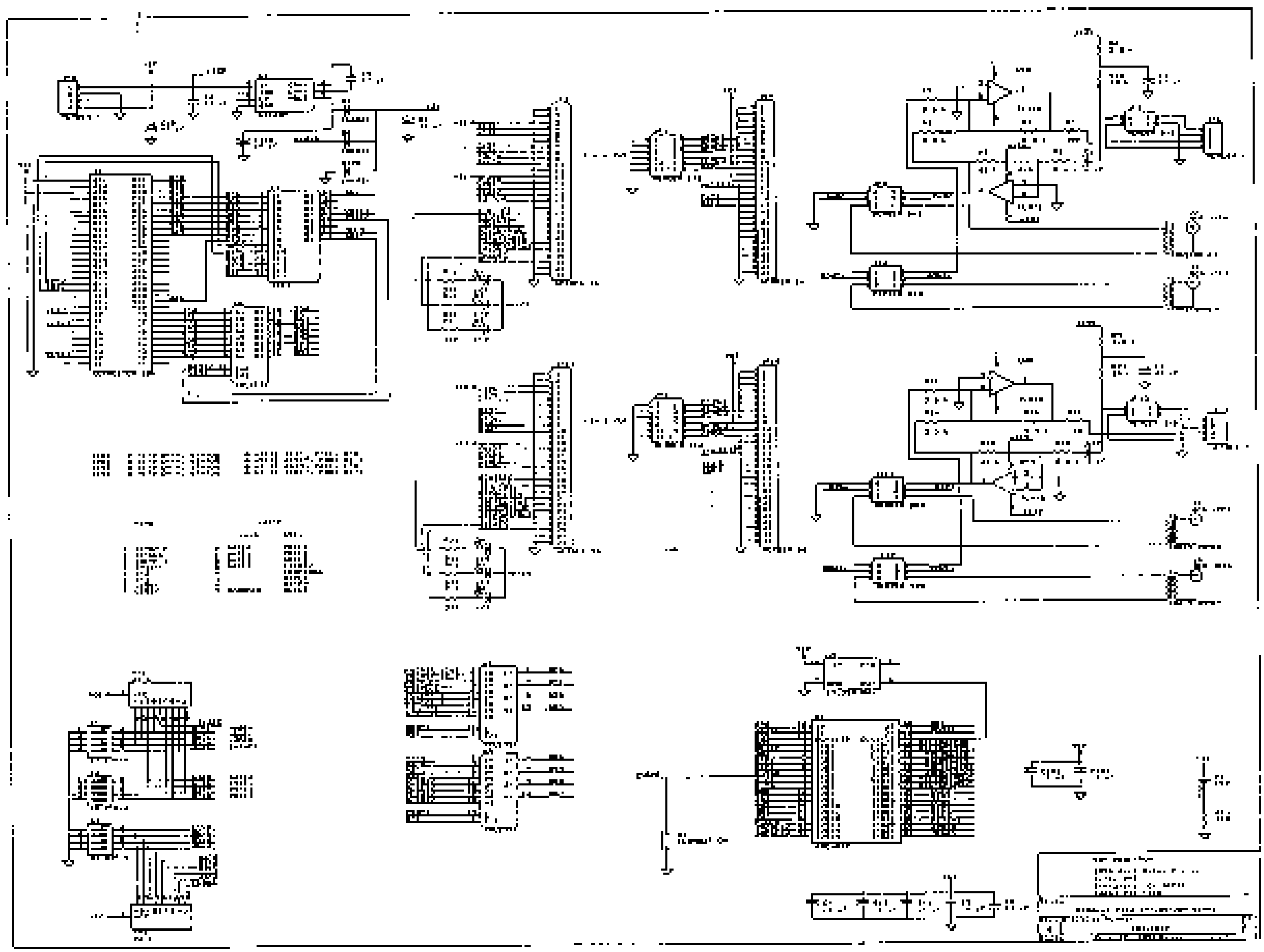


Figure A-1. Piranha 3111 evaluation board shematic (from P3111 reference manual).

We also wish to evaluate the vocoders themselves. DSP Software Engineering, Inc., of Bedford, MA, wrote the vocoder software included with the P3111. A detailed evaluation of this code is beyond scope. However, we will summarize discussions with DSP Software Engineering regarding their in-house validation efforts.

TEST CONFIGURATION AND VALIDATION

We configured the following items of test equipment (figure A-3):

1. Wavetek-Rockland 5100 Frequency Synthesizer.
2. Iwatsu SS-57110 Oscilloscope.
3. Ariel DSP-32C Data Acquisition and Coprocessor Board.
4. Hypercertion Hypersignal software.
5. AT-486 compatible PC.

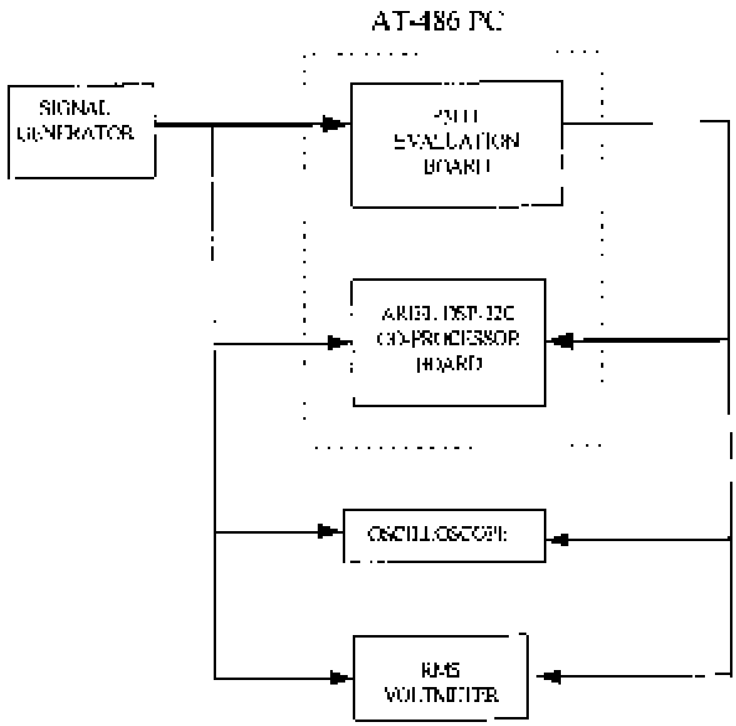


Figure A-3. Test configuration.

In a test such as this, the performance of the test equipment must exceed that of the device under test. Ariel claims true 16 bit conversion with

Signal-to-Noise Ratio (SNR)	>90 dB
Total Harmonic Distortion plus noise	0.0O05%
Dynamic Range	90 dB

No data was available for the Wavetek 5110.

We characterized the test system by applying a tone from the Wavetek 5110 directly to the analog input of the Ariel board. The spectrograms, generated by the Hyperception software, reflect distortion introduced by both the signal generator and the spectrum analyzer.

The input level to the Ariel board was 1.2 volts peak-to-peak (Vpp). This level is just below A/D saturation, as manifested by a rapid increase in harmonics and the noise floor.

While sampling at 25 kHz, the response of the Wavetek 5100- Ariel 32C combination is flat within ± 0.5 dB up to the 3-dB cutoff frequency of 11.7 kHz. Image frequencies are well attenuated; all images arising from signals beyond 14 kHz are below the noise floor of spectrograms discussed below.

Figures A-4 through A-8 show spectrograms for input tone frequencies of 0.1, 1.0, 2.0, 3.0 and 4.0 kHz. Note the Spur-Free Dynamic Range of about 70 dB. The test system SNR is estimated as (reference 1.):

SNR $\sim=$ -Noise Floor Estimate (in dB below tone power)

- 10*log(FFT length/Equivalent Noise Bandwidth).

Since	Noise Floor	$\sim=$	–110 dB,
	FFT length	-	2048 samples,
	Equivalent Noise Bandwidth (Blackman Window)	-	1.73 bins,

SNR $\sim=$ 110 dB - 10*log(2048/l.73)

$\sim=$ 79 dB.

P3111 EVALUATION

The installed PROM causes the P3111 to sample at 8.0 kHz. The observed 3-dB bandpass is 20 Hz to 3.7 KHz and is flat within ± 0.2 dB as measured by an RMS voltmeter at selected frequencies.

Figures A-9 through A-12 show spectrograms for input frequencies of 0.1, 1.0, 2.0 and 3.0 kHz. The spectrum analyzer sample rate is 25 kHz, while the P3111 A/D and D/A sample rate is 8 kHz. Spur and harmonics are below –70 dB. This is consistent with the TI TLC32044C data sheets (figures A17 & A18) which indicate second and third order harmonic distortion levels below –70 dB for input levels within 20 dB of the reference level.

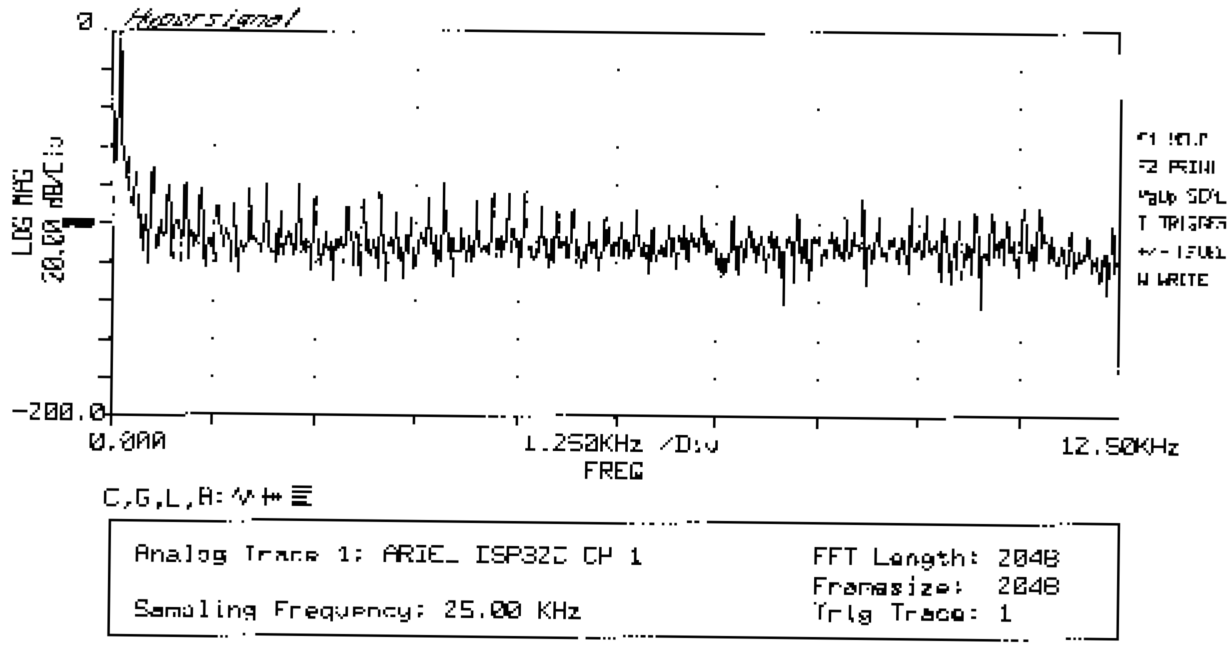
The SNR is calculated, as above, as

SNR $\sim=$ 100 dB - l0*log(2048/1.73)

$\sim=$ 69 dB.

The spectrograms of figures A-9 through A-16 are for input frequencies of 3.5, 3.9, 4.1, and 4.5 kHz. These reflect images produced by both the A/D and D/A. Given an input frequency near but below the Nyquist Frequency of Fs/2, Fin = Fs/2 – Delta, the D/A produces an image at Fs/2 + Delta, which is attenuated by the D/A output low pass (interpolating) filter.

Similarly, an input frequency, Fin = Fs/2 + Delta, is attenuated by the A/D input band-pass filter and imaged to Fs/2 – Delta. This in turn is imaged to Fs/2 + delta by the D/A and attenuated by the D/A filter output filter.



Spectrum analyzer sample rate = 25 kHz.
P3111 sample rate = 8 kHz.

Figure A-4. Test spectrum spectrogram, input frequency = 0.1 kHz.

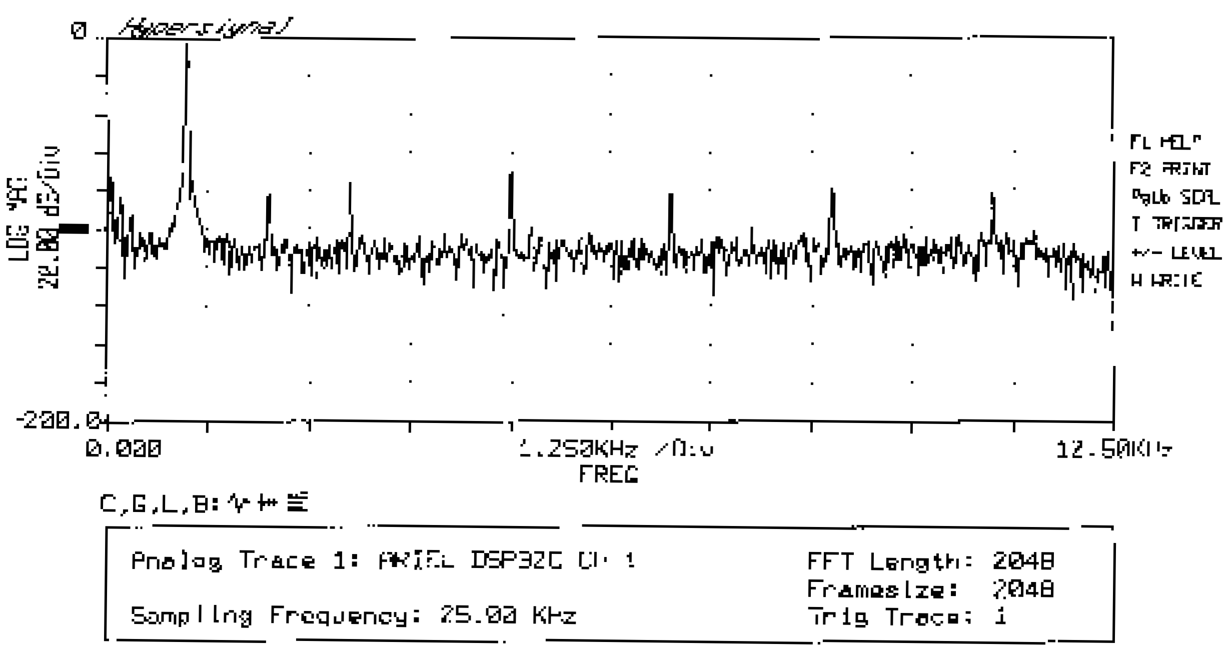


Figure A-5. Test spectrum spectrogram, input frequency = 1.0 kHz.

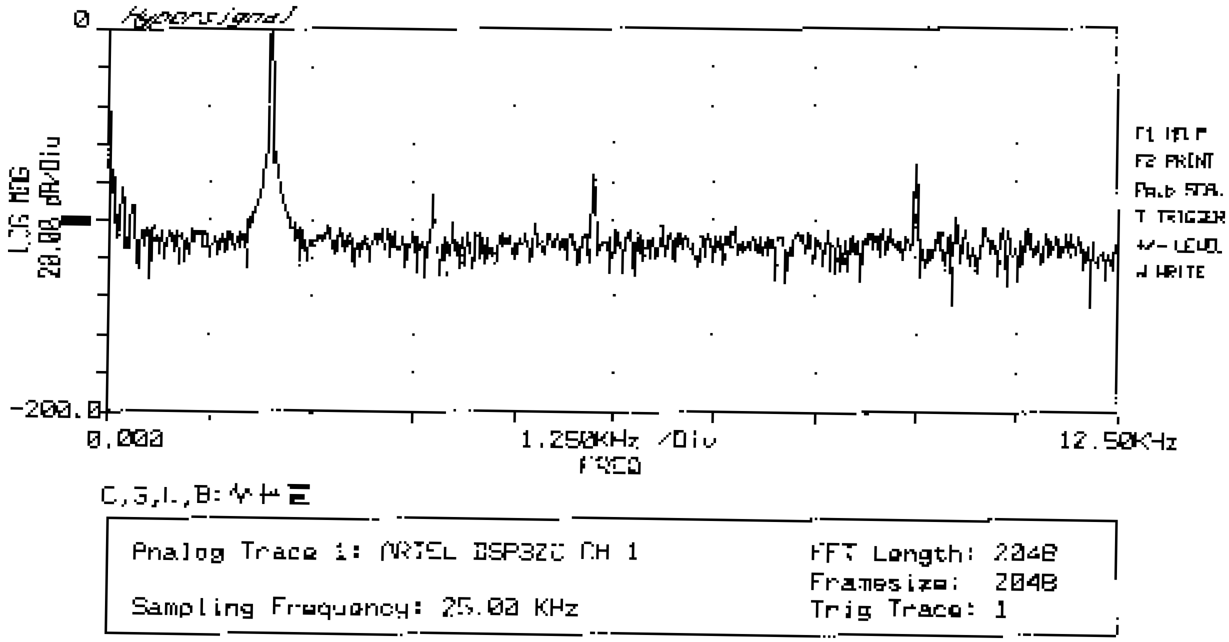


Figure A-6. Test spectrum spectrogram, input frequency = 2.0 kHz.

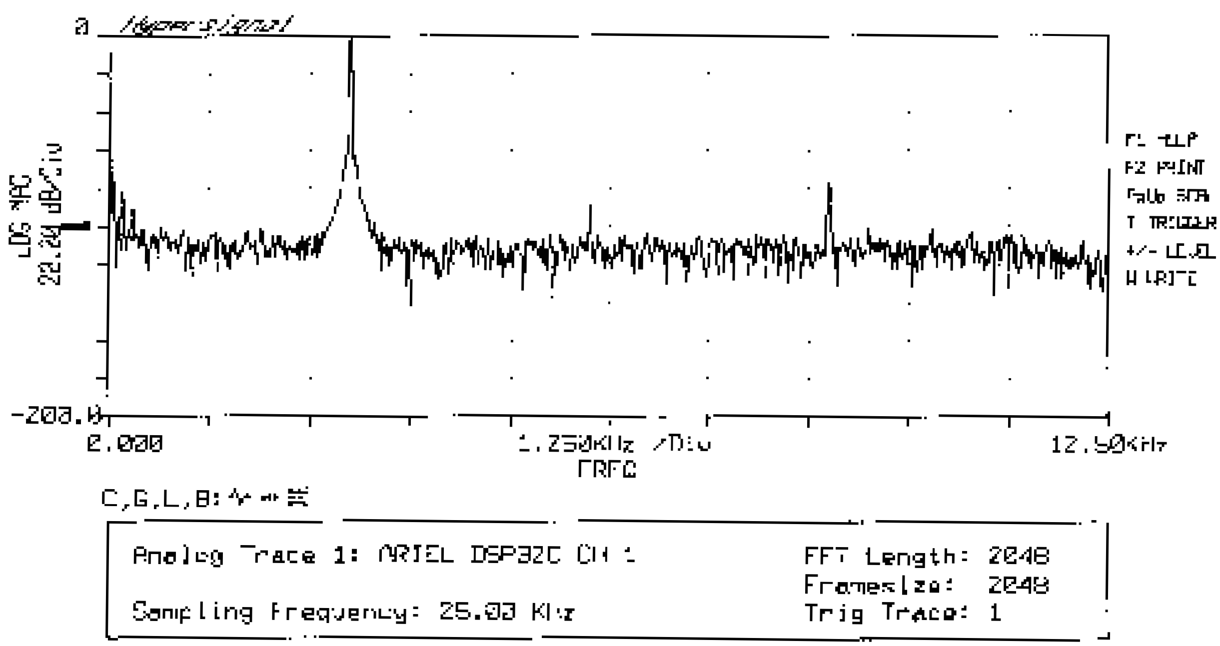


Figure A-7. Test spectrum spectrogram, input frequency = 3.0 kHz.

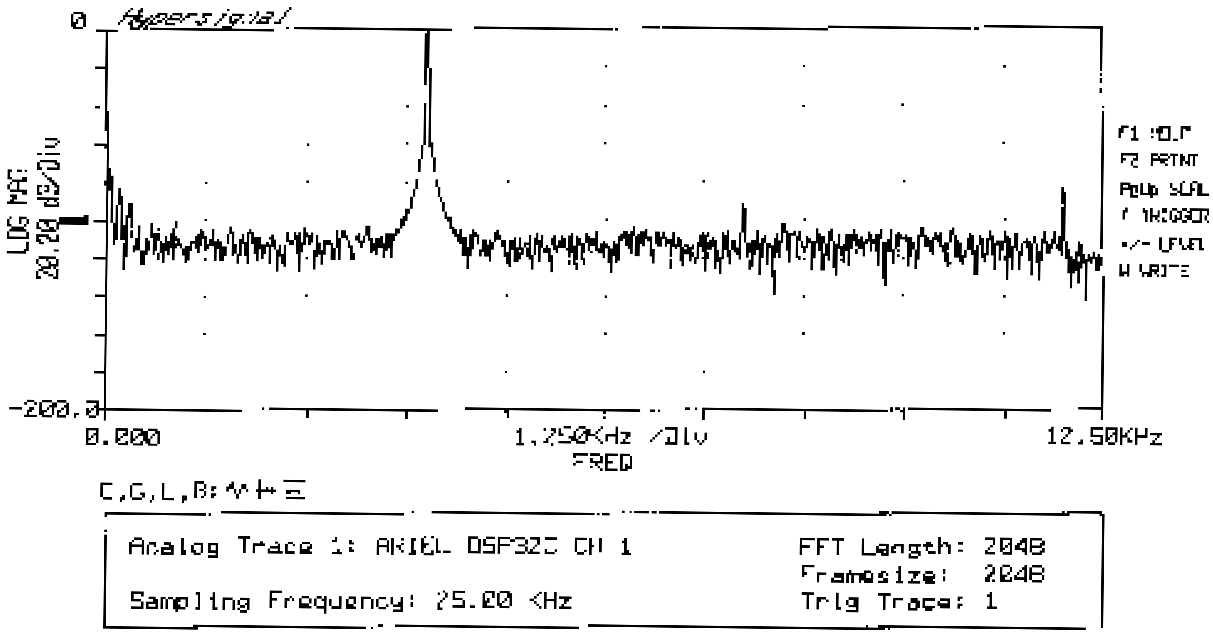
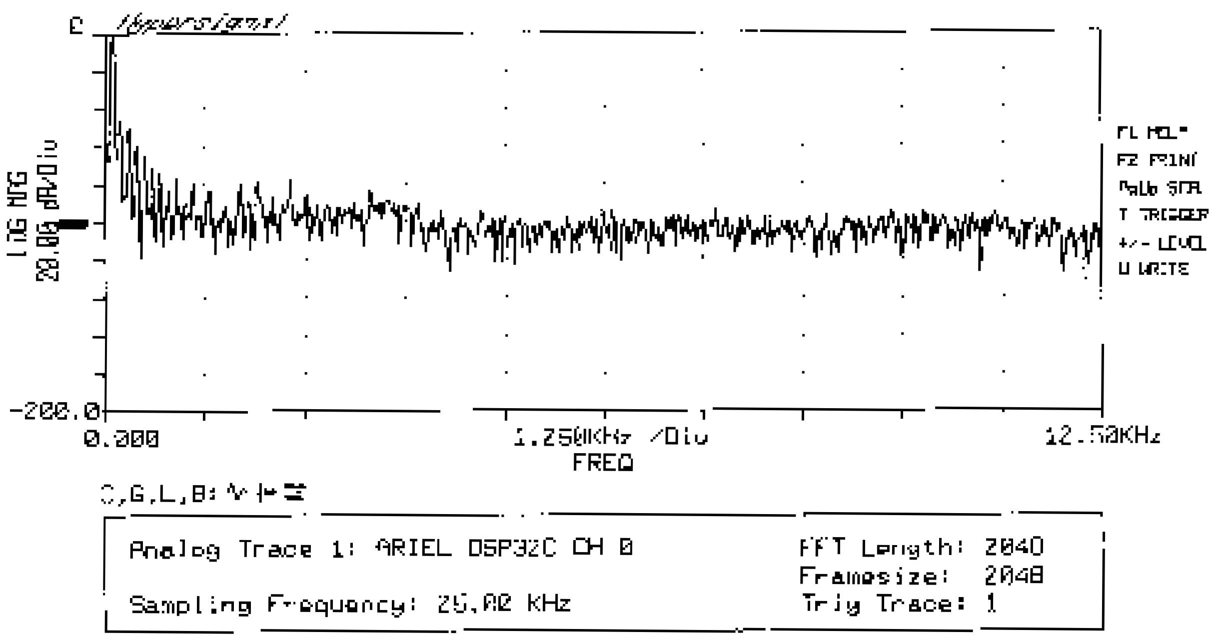


Figure A-8. Test spectrum spectrogram, input frequency = 4.0 kHz.



Spectrum analyzer sample rate = 25 kHz.
P3111 sample rate = 8 kHz.

Figure A-9. Spectrogram of P3111 output, input frequency = 0.1 kHz.

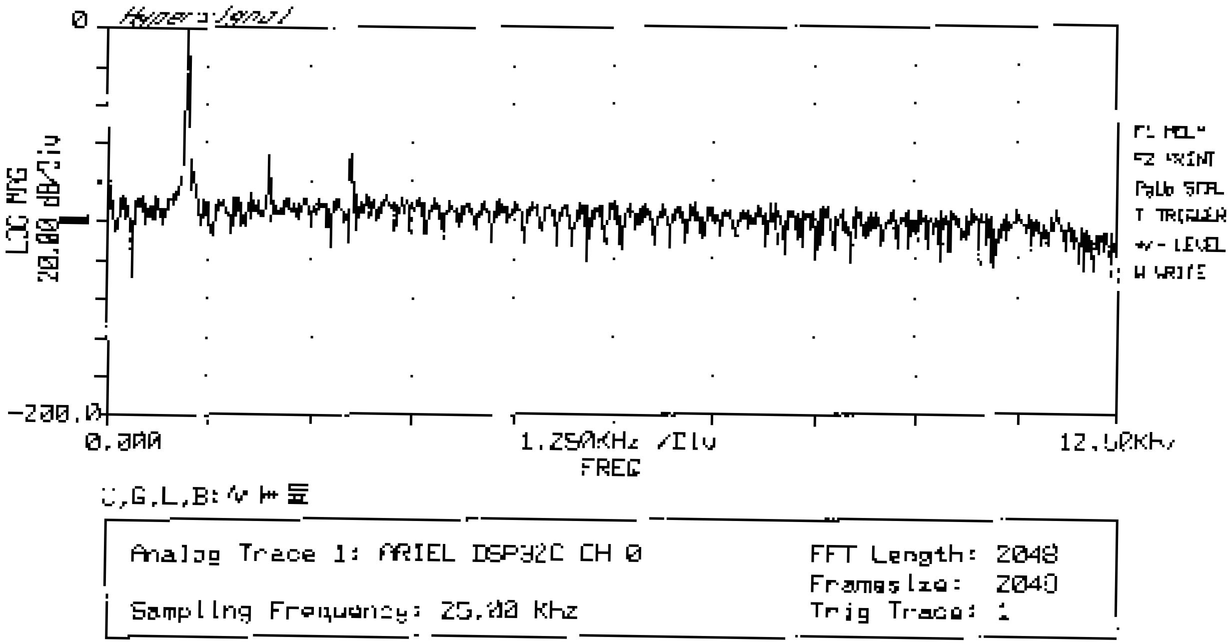


Figure A-10. Spectrogram of P3111 output, input frequency = 1.0 kHz.

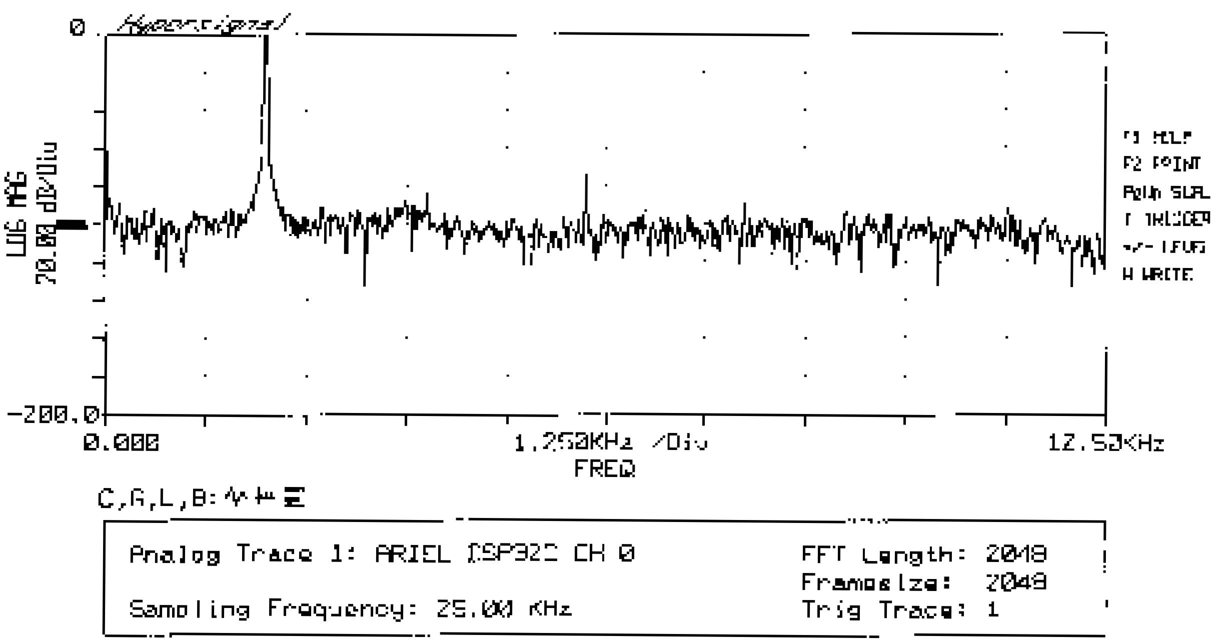


Figure A-11. Spectrogram of P3111 output, input frequency = 2.0 kHz.

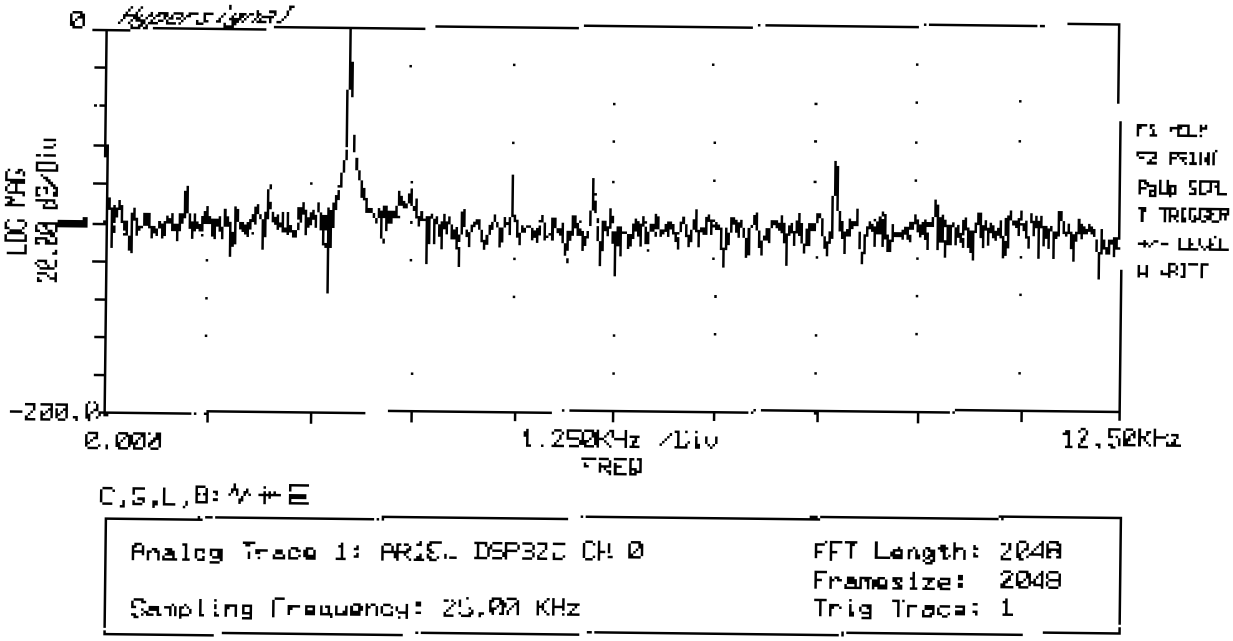


Figure A-12. Spectrogram of P3111 output, input frequency = 3.0 kHz.

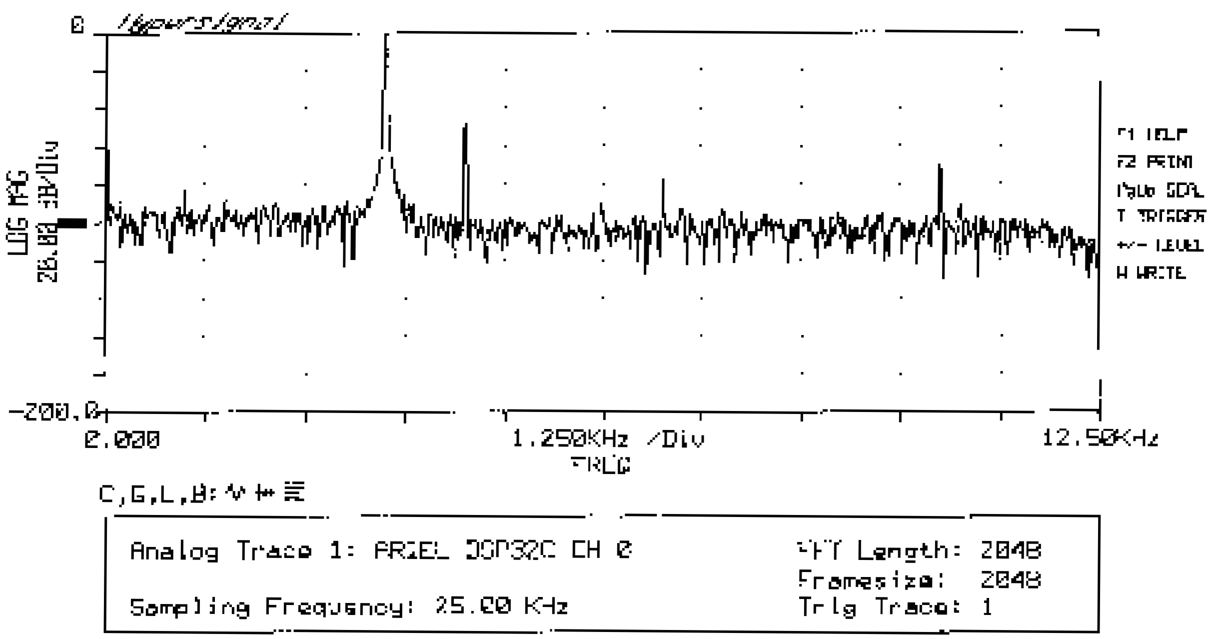


Figure A-13. Spectrogram of P3111 output, input frequency = 3.5 kHz.

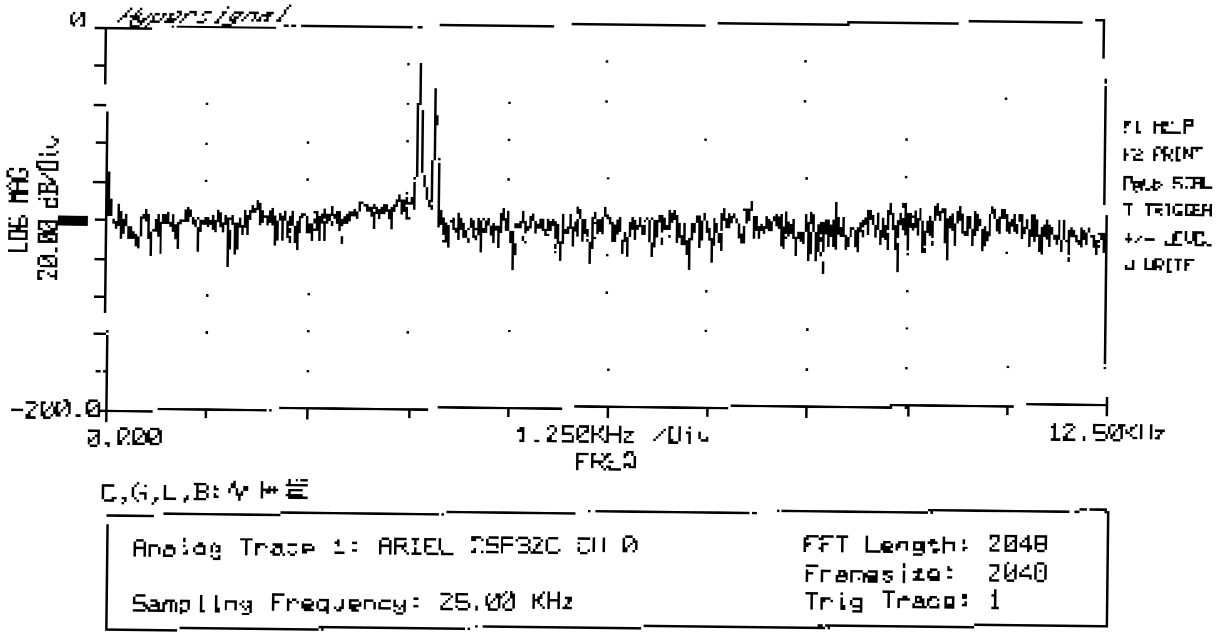


Figure A-14. Spectrogram of P3111 output, input frequency = 3.9 kHz.

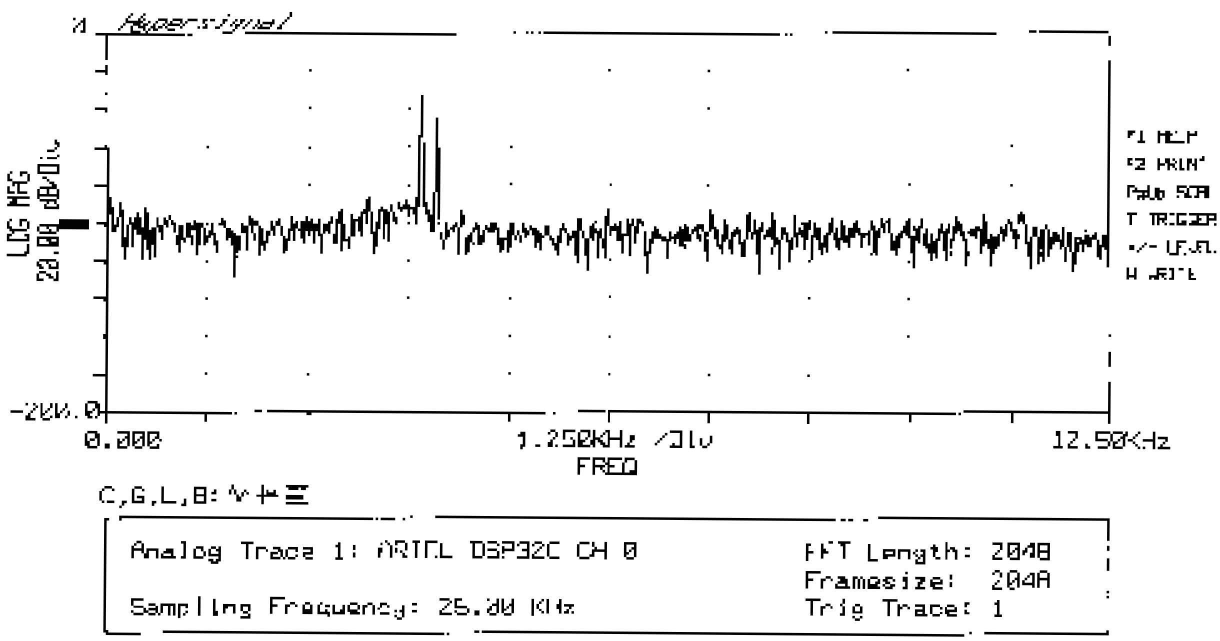


Figure A-15. Spectrogram of P3111 output, input frequency = 4.1 kHz.

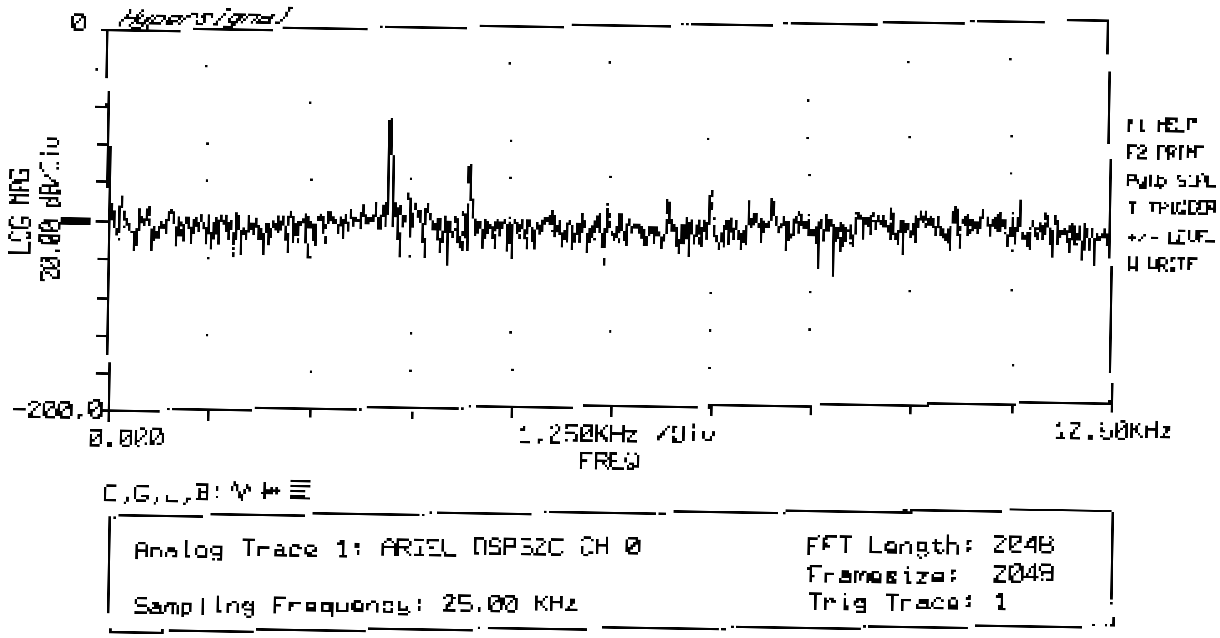


Figure A-16. Spectrogram of P3111 output, input frequency = 4.5 kHz.

The image levels in figures A-17 through A-18 generally match the filter responses from the TI data sheet (figure A-19). The filter skirts shown in figure A-19 are steep, about 120 db/octave. Despite these sharp filters, some distortion can be expected for input signals near the Nyquist Frequency of 4.0 kHz.

In summary, the analog performance of the P3111, in its evaluation board, is set primarily by the TI TLC32044 Voice Band Analog Interface chip. The designers of the TLC32044 have realized much of the potential of the 14-bit on-board converters. Estimated performance measures are summarized in table A-2.

Table A-2. P3111 performance evaluation summary.

Harmonic Levels	< –65 dB
SNR	> 79 dB
Maximum Input Level	3.2 Vpp
Voltage Gain	1.4
Pass Band (3 dB)	20 Hz to 3.7 kHz
Pass Band Ripple	±0.2 dB

VOCODER VALIDATION

Vocoders of interest here are in table A3. For each of the vocoders implemented on the P3111, a set of test vectors and a validation procedure is available from the controlling organization. DSP Software Engineering states that they have tested and validated each of the P3111 vocoders in accordance with these procedures.

TYPICAL CHARACTERISTICS

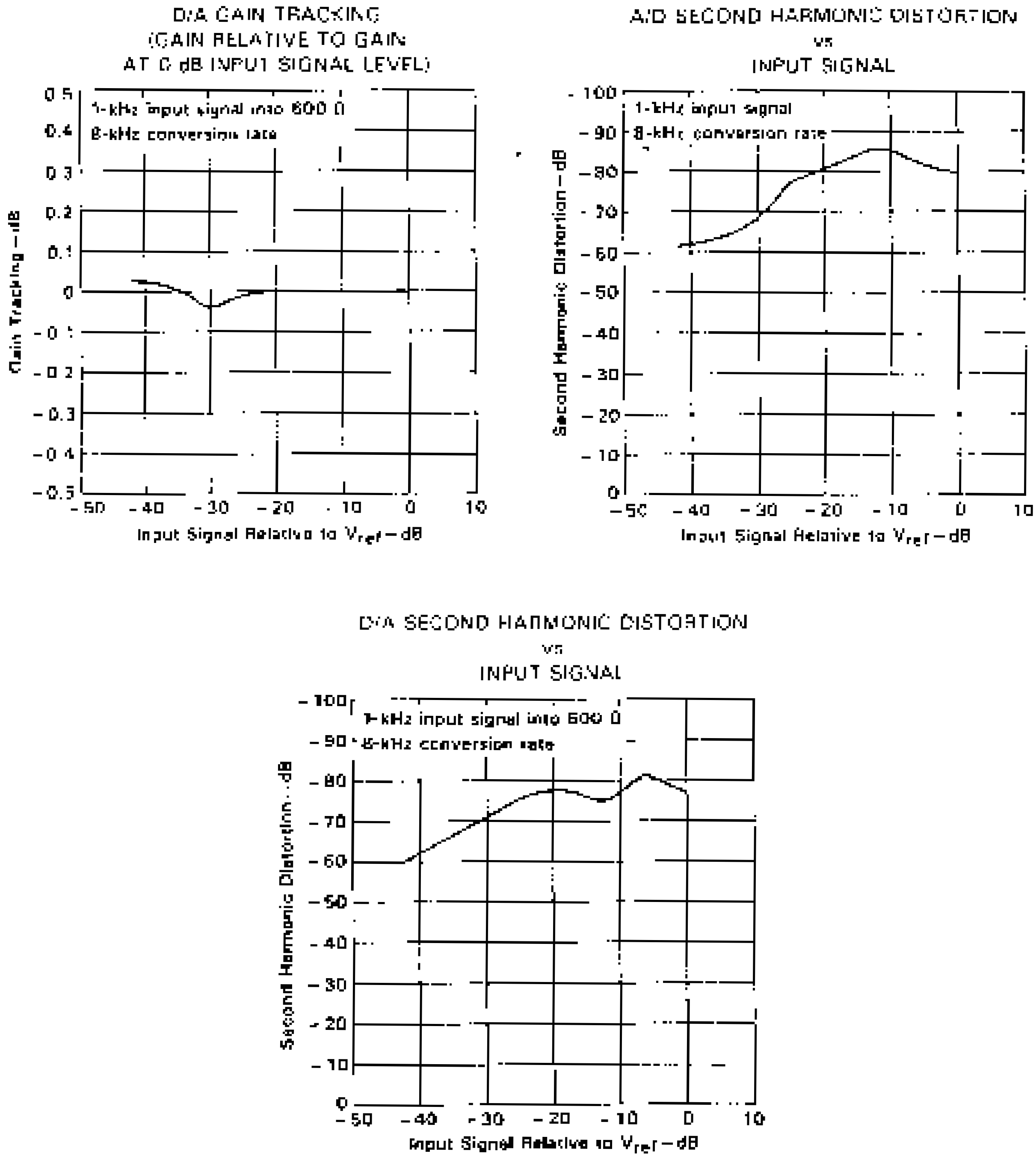
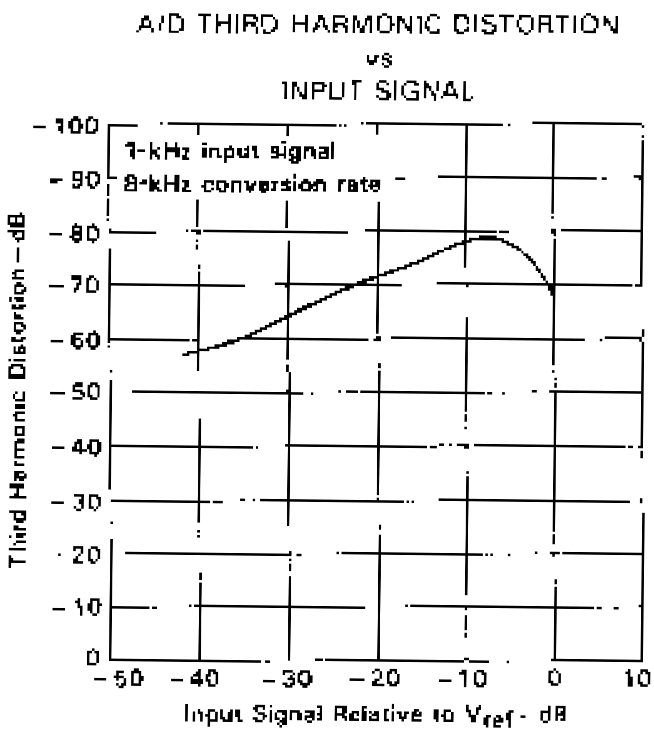


Figure A-17. TLC32044 A/D and D/A second order harmonic distortion.

TYPICAL CHARACTERISTICS



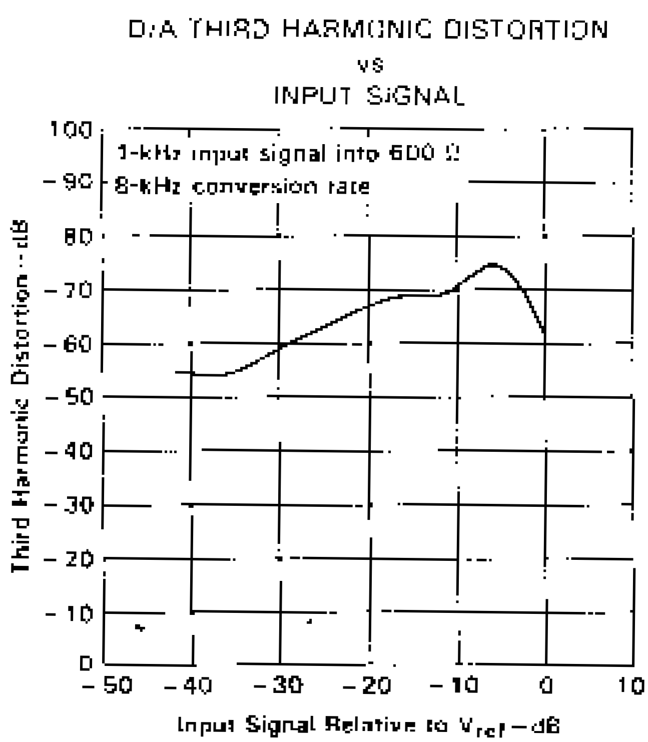


Figure A-18. TLC32044 A/D and D/A third order harmonic distortion.

TYPICAL CHARACTERISTICS

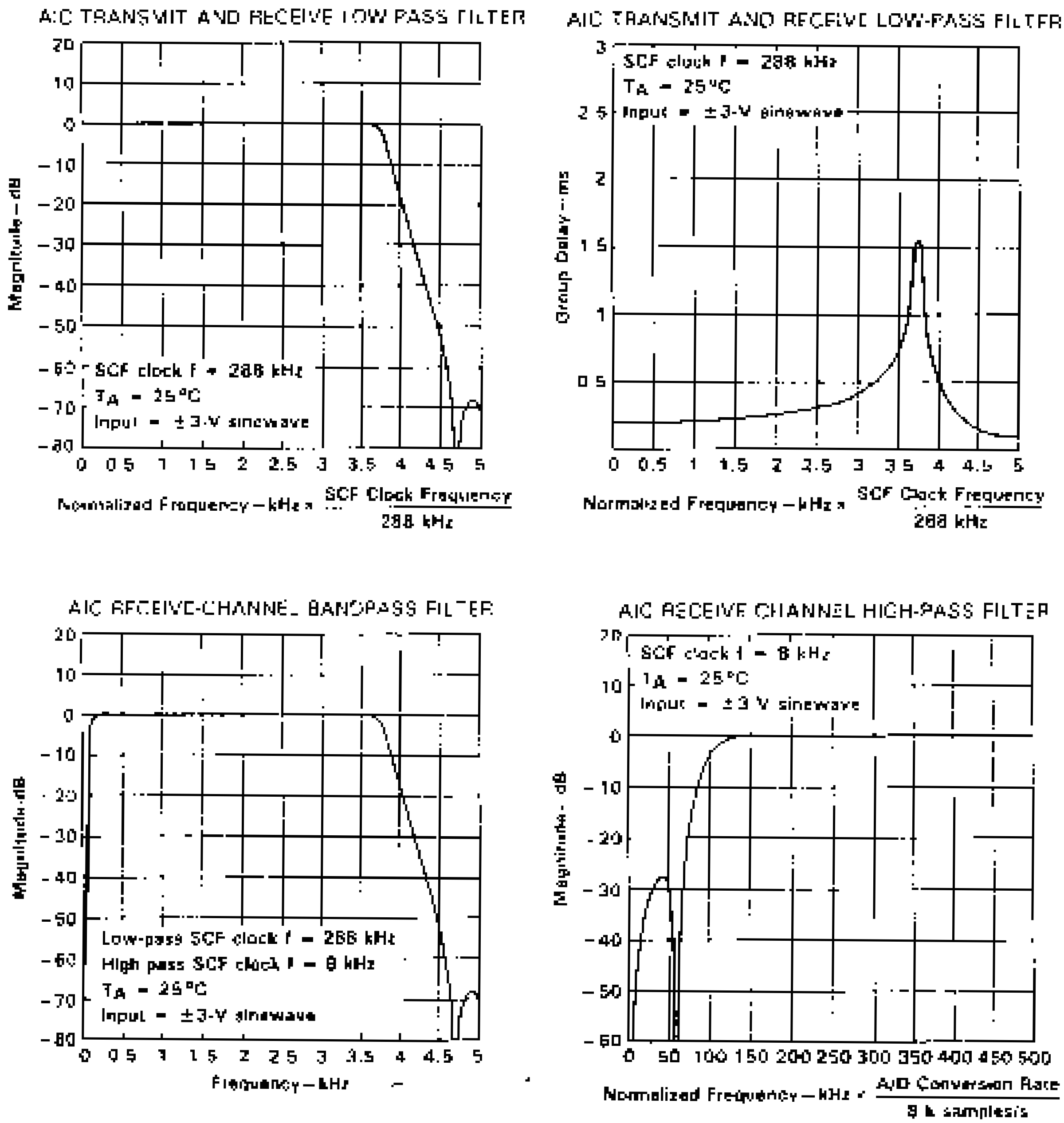


Figure A-19. TLC32044 filter response.

Table A-3. Vocoders of interest.

LPC10e (2400 bps)	U.S. Federal Standard 1015 (Linear Predictive Coder – 10 coefficients). Developed by the Department of Defense (DoD) and AT&T. Used for STU-II telephones.
CELP (4800 bps)	U.S. Federal Standard 10106 (Code Excited Linear Predictive). Developed by DoD and AT&T. Used for STU-III telephones.
VSELP (8000 bps)	The U.S. Digital Cellular Standard, TIA IS–54 (Vector Sum Excited Linear Predictive). Developed by Motorola.
LD–CELP (16000)	ITU G.728 standard. Low Delay CELP. Developed by AT&T. Used for video conferencing applications.

For all the vocoders of table A3, except LD-CELP (G.728), the verification procedure requires bit-for-bit compatibility with a reference sequence. This is not possible for LD-CELP since this is a floating point standard. Here the validation procedure requires that the deviation between the test output and a reference coder be sufficiently small. Reference 2 describes validation of the LD-CELP vocoder performed by DSP Research.

The first three vocoders in table A2 provide for a synchronization bit in their frame structure; this bit toggles every n frames. This is called “in-band” synchronization.

The G.728 standard does not allow for a sync bit. Here DSP Software Engineering has chosen to steal a bit every 16 subframes for synchronization. The bit used is the MSB of a code vector index. Thus, during the 16th subframe, only the upper half of a code book is searched. The resulting performance is minimal.

REFERENCES

1. Harris, F. J., “On the Use on Windows for Harmonic Analysis with the Discrete Fourier Transform,” Proceedings of the IEEE, Vol. 66, No. 1, pp. 51-83, January 1978.

1. Doty, D. D., “A Real-Time Implementation of the CCITT G.728 16000 BPS Low-Delay CELP Voice Coding Algorithm on a TMS320C31.” Available from DSP Software Engineering, Bedford, MA.

1. Klein, P. M., “Advanced System Software Architecture for a Multi-Algorithm Rate Adaptive Audio/Voice Codec.” Available from DSP Software Engineering, Bedford, MA.

1. “Voice-band Analog Interface Circuits, TLC32044I, TLC32044C” (data sheet). Texas Instruments Inc.

ACKNOWLEDGEMENTS

We thank Paul Klein of DSP Software Engineering and Yang Bin Wang of DSP Research. Paul provided important details on vocoder implementations. Yang Bin developed and made available, at no cost, the loop-back PROM used in these tests.

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704-0188*

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE October 1994	3. REPORT TYPE AND DATES COVERED Final: February 1994 to August 1994
4. TITLE AND SUBTITLE VOCODED KING DATA BASE		5. FUNDING NUMBERS AN ICSX4600 PE 0405088G
6. AUTHOR(S) P. C. Grossnickle		
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Command, Control and Ocean Surveillance Center (NCCOSC) RDT&E Division San Diego, CA 92152–5000		8. PERFORMING ORGANIZATION REPORT NUMBER TD 2721
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) National Security Agency Central Security Service 9800 Savage Road Ft. George G. Meade, MD 20755		10. SPONSORING/MONITORING AGENCY REPORT NUMBER
11. SUPPLEMENTARY NOTES		
12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.		12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This document reports the development of a data base of vocoded speech that has been passed through several popular voice coders and their corresponding decoders. The KING data base consists of samples from 26 male speakers recorded sequentially on DAT tape. Two products were produced: five DAT tapes, each consisting of the original KING data base on one channel and the vocoded version on the other; and raw binary files, one for each vocoded and speaker combination.

14. SUBJECT TERMS vocoded speech vocoders decoders			15. NUMBER OF PAGES 16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT SAME AS REPORT